MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer

Year: 2024 | Citations: 135 | Authors: Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo

Abstract

The recent large-scale text-to-speech (TTS) systems are usually grouped as autoregressive and non-autoregressive systems. The autoregressive systems implicitly model duration but exhibit certain deficiencies in robustness and lack of duration controllability. Non-autoregressive systems require explicit alignment information between text and speech during training and predict durations for linguistic units (e.g. phone), which may compromise their naturalness. In this paper, we introduce Masked Generative Codec Transformer (MaskGCT), a fully non-autoregressive TTS model that eliminates the need for explicit alignment information between text and speech supervision, as well as phone-level duration prediction. MaskGCT is a two-stage model: in the first stage, the model uses text to predict semantic tokens extracted from a speech self-supervised learning (SSL) model, and in the second stage, the model predicts acoustic tokens conditioned on these semantic tokens. MaskGCT follows the mask-and-predict learning paradigm. During training, MaskGCT learns to predict masked semantic or acoustic tokens based on given conditions and prompts. During inference, the model generates tokens of a specified length in a parallel manner. Experiments with 100K hours of in-the-wild speech demonstrate that MaskGCT outperforms the current state-of-the-art zero-shot TTS systems in terms of quality, similarity, and intelligibility. Audio samples are available at https://maskgct.github.io/. We release our code and model checkpoints at https://github.com/open-mmlab/Amphion/blob/main/models/tts/maskgct.